Advancements In Human Action Recognition Research In Last Decade: Survey Paper

Manisha Mudgal[1]

PHD Scholar, J C Bose YMCA University,FBD, HR,India, mudgal.05.manisha@gmail.com

Dr Deepika Punj[2]

Assistant Professor, , J C Bose YMCA University ,FBD,HR, deepikapunj@gmail.com

Dr Anuradha Pillai[3]

Associate Professor, , J C Bose YMCA University ,FBD,HR, anuangra@yahoo.com

ABSTRACT:

Human Action Recognition has got many applications in the field of Computer Vision. Apart from surveillance systems, Human Action Recognition is widely being used in the fields of robotics, healthcare, behavior analysis, social networking, education, and video modeling as well. In the last decade, a lot of researchers have proposed new algorithms for the development of HAR systems. This survey paper shows how HAR methods have changed in the last years. Researchers in the early years were using machine learning methods and in last recent years, deep learning gained popularity. There is a need to keep a record of all these developments. These tracks can act as a guide to new researchers. This survey aims to provide them with all the methods that new research can use to solve the problem and will help to show a path for future research. This survey covers all the significant techniques used in the past decade along with it different feature extractors and datasets used are also covered. Models used have been elaborated in detail along its future work scope has also been sighted.

Keywords: Human Action Recognition Systems, Machine Learning, Deep Learning, Classification, Datasets, CNN

1. INTRODUCTION

Computer Vision research in the field of Human Action Recognition has gained a lot of attention in past decades. Human Action Recognition has varied applications which have made many researchers throughout the world work on it. In HAR actions of one or more people are detected and recognized by the system. It can be applied in many areas like healthcare, robotics, retrieval of video content, Video Surveillance, scene modeling, etc. It can be installed in open areas to monitor anonymous activities like parks, parking, airports, etc. These HAR systems also play a very crucial role in observing patients with motor disabilities and mental conditions. These live assisting systems can ease the monitoring task and assure safety.

Action Recognition Systems uses the process of recognizing the visual movements and classifying the actions or movements by naming the action that best suits the instance. There can be complex videos also where multiple people are performing different actions at the same time. The action Recognition system must be capable of classifying multiple actions at the same time also. Sometimes background objects and things also provide extra information about the event.

The basic steps that are followed by any HAR system are:

- Preprocessing the data
- Extracting Features
- Learning Actions
- Action Classification

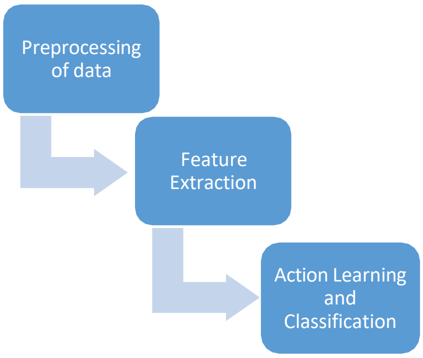

Fig 1: Basic steps of HAR system

PreProcessing involves the steps used for cleaning, smoothing, and grouping the data. In preprocessing, data cleaning is used to remove inappropriate data. Smoothing involves removing noise and grouping includes methods to establishan association between variables.

Feature extraction is used to get the best features from the data sets. These features are extracted using selection and combination of variables into features. These features are easy to process and still have the ability to describe thedata with originality. This task is sometimes challenging due to noise, improper lightning, background objects.

The feature Extraction process can future be classified into different classes:
Shaped-Based Feature Extraction: Shape-based is measuring similarities between shapes that are represented by features. It uses two methods feature extraction and similarity measurement between extracted features.
Spatio Temporal: This method discovers useful patterns from the data collected over time and space.
Optical Flow-Based Method: It represents the visual features like edges, corners, interest points. Histogram of Optical Flow uses information of Optical Flow to describe normal patterns.

Interest Points: Spatio Temporal Interest Point based detector captures the IP from spatio temporal domain.

In Action recognition, new models are trained on extracted features and recognize the features according to t the class. After that testing on datasets is done by performing classification.

In the Machine Learning approach, there is need to specify and adjust the feature extractor and classifier. In the deep learning approach, human intervention is the least. This method will automatically find out the features.

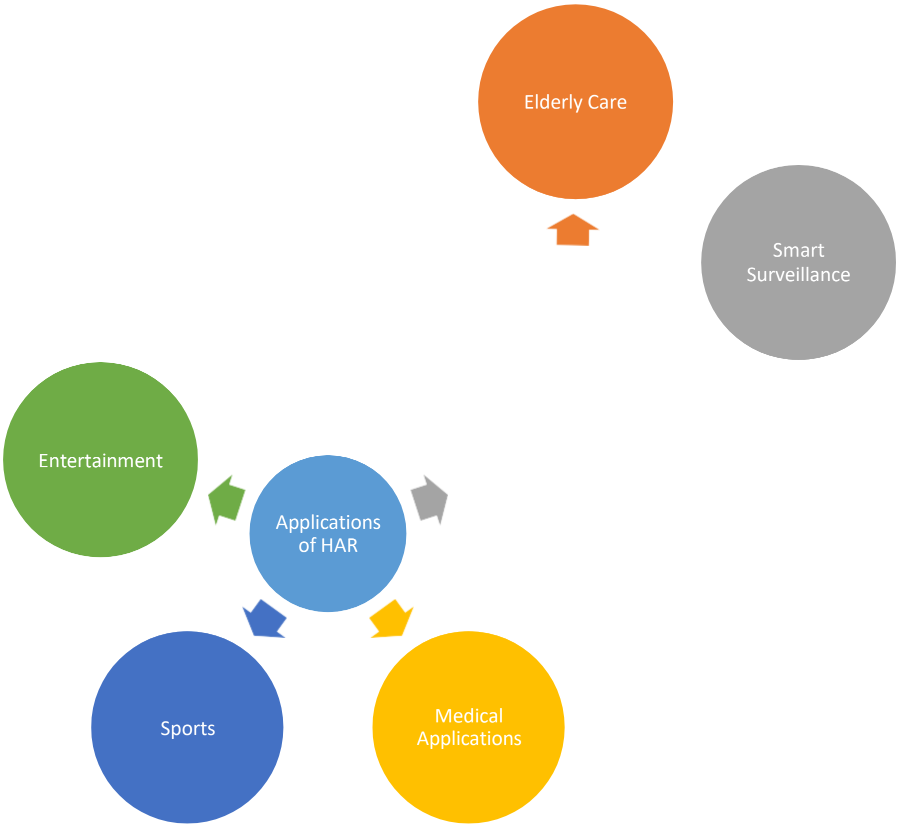

Fig 2. Showing some applications of HAR systems

List of the things included in this paper are:

- It includes all the major databases available for training.
- It includes all the approaches machine learning approaches.
- Then all deep learning approaches have been included.
- Finally, discussion on various prospects has been included for those interested in future research.
-

2. Datasets Available For HAR

Datasets used in this system are mainly video-based and still image-based. Video-based datasets contain a variety of action clips and image based dataset contains stills. Some of the datasets are discussed below:

2.1 Image Dataset

2.1.1 Pascal Dataset

The Pascal dataset contains a collection of still images. This dataset contains classes related to jumping, calling, riding, walking, reading etc. Initially, the dataset was of 4 classes and in 2007 it reached 20 classes.
The dataset is divided into training and testing with the ratio of 2:1 or 1:1 as per the choice of the user.

2.1.2 Stanford 40 Actions

This Dataset contains about 40 action categories or classes performed by Human beings. Ihere are around 9532 images in total. Out of which 180 to 300 images are related to one action. B. Yao et al. [1] in November 2011. have also usedthis dataset in their paper.

There are some other datasets also which have still images related to human activities like People Interacting with Musical Instruments (PPMI), Sports Dataset –which contains images of 6 different sports.

2.1.3 WILLOW ACTION DATASET

This dataset contains images related to 7 different classes. Each class contains around 108 images. Seventy percent of the data is used to train the model and remaining of the data is used to test the modal.
Someone some researchers have used these images with the bag of features method. One of the researchers Liag et al.[2] has used it with a novel deep learning model and gained accuracy of almost 80.4%.

Dataset Name	Year	Reference	Accuracy
Pascal	2017	Zhao et al.[3]	93.4%
Stanford 40 Actions	2017	Zhao et al.[3]	91%
WILLOW ACTION	2015	Liang et al.[2]	80.4%

Table 1 Showing Results of HAR systems using Still Image Datasets.

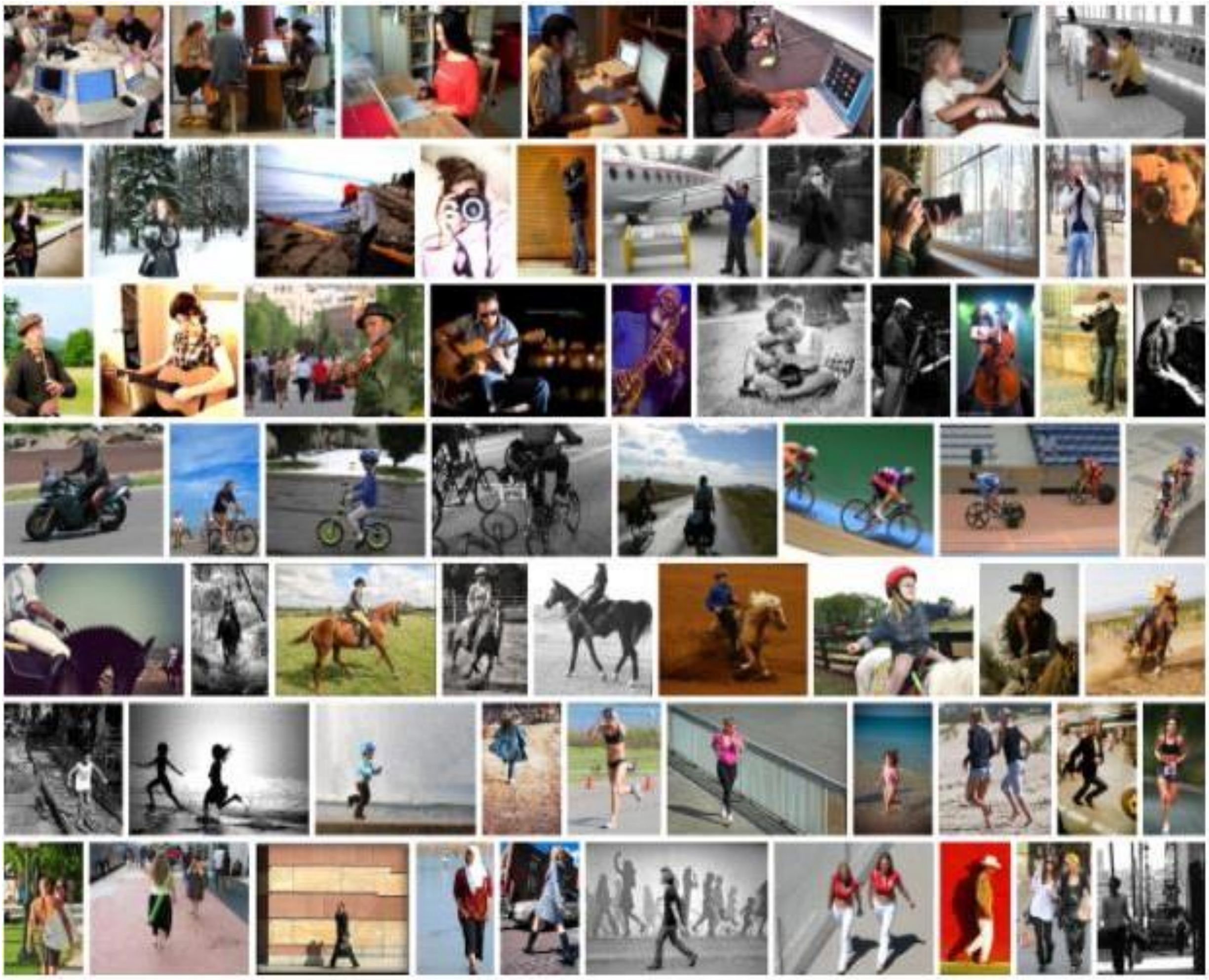

Figure 3. Example of Willow Action Dataset showing interaction of human with Computer, Guitar, Camera and some other activities like horse riding, bike riding, running etc. [57]

2.2Video Datasets

2.2.1 UCF 101 Dataset

It is an action recognition dataset where videos are based on realistic actions and are collected from Youtube. There are around 101 action categories. Around 13320 videos are there for 101 action categories. Action Categories are namely like jumping ack, Typing, Jump Rope, Hammering, Head Massage, cutting in the kitchen, Paying Piano, Pull-ups, Push-Ups, Soccer, Sky Diving, Long jumps, Cliff Diving, Gymnastics, Fencing, Bowling, etc

2.2.2 KTH Dataset

This consists of 6 types of human activities like clapping, jogging, walking, waving, boxing, running. There are around 600 videos in a total of all the actions. These videos are used with various machine learning and deep learning approaches.

Many reachers have used this dataset in their projects and have achieved accuracy up to 90.3 %.

2.2.3 WEIZMANN Dataset

The Weizmann dataset contains videos related to 10 actions like jumping, running, bending, skipping, jumping jack, walking, jumping in place, two hands waving, one hand waving. There are around 90 videos and each of these activities is done by 20 different human beings. Viewpoints is same and therefore it does not work good in many scenarios of real life. There are 25 frames per video with a resolution of 140*180 pixels.

2.2.4 HOLLYWOOD2 DATASET

The Hollywood 2 dataset contains 12 action categories with 2517 videos and these gives more real life situation challenges as compared to other datasets. Some of the actions included are fighting, hugging, exiting the car, sit up, eating, and shaking, talking on the phone, standing up etc.

There is a lot of variations in the videos that suit the real life scenarios. There are videos in different postures, motions,camera motions, occultation's, illuminations, etc.

Many researchers have used this dataset for their model and Islam N et al. [4] has shown accuracy up to 87.2%.

Dataset Name	Year	Reference	Accuracy
UCF-101	2019	Ullah et al.[6]	94.33
KTH	2019	Sharif et al.[8]	99.9
HOLLY WOOD2	2019	Islam et al.[5]	87.2
WVU	2018	Li et al.[7]	Results from all 8 viewsis average

Table 2 Showing Results of HAR systems using Video Datasets.

DATASET	No. of classes	Link
PASCAL VOC	20	http:// host. robots. ox. ac. uk/ pascal/ VOC/[58]
STANDFORD 40	40	http:// vision. stanf ord. edu/ Datas ets/ 40act ions. Html[59]
WILLOW ACTION	7	https:// www. di. ens. fr/ willow/ resea rch/ still actio ns/[60]

Table 3 Image Datasets with download links

3.MACHINE LEARNING ERA

In machine learning, firstly data is collected, then features are extracted, and then the extracted features are provided to the learning model. HAR systems had mostly used techniques of ML initially. Vision-based works like image representation and classification were carried out. Earlier silhouettes and complete images were used to represent actions globally. HOOF (Histogram of Optical Flow) and HOG (Histogram of Oriented Gradient) were used for localrepresentation. There are several classifiers like SVM, KNN.

HRA Classification

HRA can be classified into two basic categories: Unimodal and Multimodal. These can be future classified. The single modality datasets is used by Unimodal to identify actions. This modal can future be classified as shown in the diagram below.

Spatio Temporal method:

Spatio Temporal is the most commonly used technique of Unimodal. It is used to analyze the movements of humans in detail. This uses 3D body representation for the localization of actions. This can also show sensitivity to noise andcomplex data.

It has shown very impressive results for simple datasets like KTH and Weizmann (accuracy of around 99.3%) but complex dataset's results were not impressive.It has shown very impressive results for simple datasets like KTH and Weizmann (accuracy of around 99.3%) but complex dataset's results were not impressive.

SIFT and HOG-based methods shifted to histogram-based on optical flows and were called Histogram of Optical Flow HOOF. HOOF does not need any background subtraction or pre-processing. The features are independent of direction. In 2009 Gall et al.[17] achieved around 4.4% accuracy using HOOF on Weizmann Dataset.

Shabani et al.[18] used a temporal filter for asymmetric synchronization as it gives a better result as compared to Gabor Filter.

Sahoo et al. used a fusion of histogram-based features to perform the recognition of action.

In 2015 Nguyen et al. proposed spatiotemporal attention pooling method. The proposed

method calculates the similarities between the backgrounds at different levels and then the results are given to SVM.
Stochastic Method

Various methods like Hidden Conditional Random Fileds[19] ,Dynamic Bayesian Netorks [20] of Stochastic methods have been developed. In 2000, Oliver et al.[21] Proposed an algorithm which can models the human's interaction. The algorithms formed a feature vector that kept track of human movement and described the motion of human.
In 2013 Prince DSJD [22]used Stochastic Markov Random Forest framework to detect and identify activities in a video. A fixed length feature vector was used to represent the action clip by Perronnin et al.[23] in fiser kernel technique.

Stochastic Method:

The stochastic method is the modern technique of Unimodal. This is shown usefulness for complex datasets. The problems of occlusion, cluttering have been dealt with efficiently using this method. It has shown high accuracy on complex datasets. Stochastic methods have also some cons like it uses label bias and there can be an issue of overfitting as it needs a large amount of training data.

Various methods like Hidden Conditional Random Fileds[19] , Dynamic Bayesian Networks [20] of Stochastic methods have been developed. In 2000, Oliver et al. [21] Proposed an algorithm that can models human's interaction. The algorithms formed a feature vector that kept track of human movement and described the motion of humans.
In 2013 Prince DSJD [22] used the Stochastic Markov Random Forest framework to detect and identify activities in a video. A fixed-length feature vector was used to represent the action clip by Perronnin et al.[23] in fiser kernel technique

Rule Based Methods:

For Group or Multiple action recognition, the Rule-based method is used. This method is mainly used on short videos and a long video needs to be broken in small clips. This method got more popular in 2009 and gives more promising results on simple datasets. Generation of suitable attributes is difficult in complex datasets.
Wang and Mori in 2010[24] presented a model which captured the correlations between various attributes and increased the accuracy of recognition.

Jayaraman and Grauman 2014 [25] worked on animals with an attribute, sun scene attributes. zero-shoot training technique was used. This approach models the untrusted attributes due to the prediction of the classifier and due to possible incompatibility with unseen classes.

.Shape Based Method:

poses and silhouettes forms are used to represent body in Shape-based method. It can recognize both still images and videos actions. This approach faces a problem when tracking of the skeleton is done and when there is occlusion, illumination.

Multimodal methods:

Recently the focus of researchers has shifted on multimodal methods for activity recognition. To describe an event different types of features can help in providing useful information.

If we have audio–visual content, then applications will not only work on an analysis of audio and visual data but alsoto track and recognize the action.

These methods are classified as:

Affective Method

The base of the emotional state of a person depends upon the mapping between an individual's emotional state and the activities. With the study of Affective modeling, a person can express, recognize all the movements like hand gestures, speech, expressions.

This area of research is the combination of AI, pattern recognition, Computer Vision, psychology. One of the biggest problems with this method is the gathering of accurately annotated data.

Ratings are considered the best way to express a person's affective state. However, in the real world, it is very difficult to express the affective state as different persons react in different ways to the same situation.

Nicolaou et al. [49] has proposed a new method that is based on PCCA.

Behavioral Method

The behavior recognition system provides information about the nature and mental state of a person. Recognition of human behaviors is one of the challenging tasks. Many factors can affect human behavior like actions, mood, emotional state, companion, and place.

Vrigkas et al. [50] developed a CRF model that can recognize human behavior as friendly, aggressive, neutral.

Social Networking based Methods

Social interaction is a very special type of activity where people interacts with other ,share their feeling through posts, videos. Facebook, Youtube, Twitter, Instagram etc are some sites that affect a person's behavior.

Fu et al. in 2014[51] performed attribute-based social activity recognition. They classified normal life activities as weddings, birthdays.

4. Deep Learning ERA

DL is a subset of ML where multiple layers are there to imitate the way the human brain works.

It is widely used in research areas like pattern recognization, image processing, computer vision etc. In this model feature, extraction is done automatically. That is why it is a good option when there is lack of awareness of the domain. Algorithms of deep learning take very little time as compared to machine learning during testing.

CNN

Convolutional Neural networks can automatically learn features. Ji et al[26] used CNN for the recognization of actionsin videos. Frames of videos were used as images
and CNN was applied to recognize the activities at every frame.

Yang et al.[27] used a CNN on Opportunity DaTASET which showed better results than the four baseline methods. Ronao and Cho [56] proposed deep ConvNets using accelerometer and data of gyroscope sensor. They used it torecognize actions.

Safaei M and Foroosh [28] introduced a novel method based on CNN to predict future action. It also detects theimage's important parts like shape and region.

Banerjee et al. [29] used CNN where 4 features were encoded on a greyscale. This was used to produce a detection score and also used fuzzy fusion to generate final decision-based confidence.

DNN

DNN is a deep network as they have more layers. Deep NN can learn much more data than others. Aubry et al. [30] used OpenPose, a DNN based detection system to extract the 2D skeleton. The sequence of motion is then convertedto an RGB image. RGB Image becomes an action sequence which is then classified by the neural network to recognizeactions.

Khan et al.[31] in 2020 presented a DNN and multiview feature extractor approach where both are used to extract features separately. The results are combined to get the best features. And this was tested on almost 5 datasets like KTh, Youtube, IXMAS, UCFSports, and the results are between 93to 97%.

Dai et al.[32] resolved the visual attention ignorance problem by proposing an LSTM based on two stream attention. This recognizes actions in videos and gained an accuracy of 98.6% on UCF sports data and 96.9% on UCF sports.

RNN

Recurrent Neural networks are the algorithms used behind Apple Siri and Google Assistant. It is one of the algorithmswhich is behind the success of deep learning over the past years.

This algorithm has internal memory and are very robust and powerful. As they have a memory to learn, they can be very precise while predicting what can come next. They are best used for sequential data like speech, text, videos, series, financial data, etc.

LI et al.[33] presented a tree based on RNN that can do adaptive learning for skeleton-based HAR. This alsocategorizes the action classes. This method was used on 3 D SAR 140 dataset and the accuracy is around 89.2%.

Qi et al. [34] presented a new approach called stagNet to develop a semantic RNN. It can recognize the group as wellas individual actions.

SAE

Auto Encoder learns through hidden layers. The process of learning is called the encoding decoding process.

Gao et al. in 2019[35] proposed an algorithm that was based on Stacked Denoising AE (reduces the noise and extracts useful features) and LightGBm for action recognization. The results showed 95.9% of accuracy.

Almasklukih et al. [36] also used two SAE and also included a softmax layer. The training of first layer was donw with 561 features using autoencoder and it gave 80 features to AE.

RBM

RBM stands for Restricted Boltzmann Machine which is a generative Stochastic Artifical Neural Network. These are shallow having two layers. RBM can learn a probability distribution over a given set of data.

In 2020, Abdellaoui M et al. [37]a HAR system based on Deep Belief network which extracts features from videos and then classification is done. Accuracy of around 91.8 % is achieved on the dataset.

Using RBM deep Belief approach Foggia et al.[38] obtained accuracy of about 85.8 on MHAD and 84.7 % on MIVIA datasets.

Hybrid Network

Hybrid models are formed by combining two or more techniques for solving a single problem. Researchers have applied a different combination of existing models with proper customization.

Kanjo et al. [39] have compared different architectures on a dataset having 40 data files. The results are like 72.9%- MLP model, 78.6% - CNN model, 94.7 % - CNN LSTMmodel.

Jaoudi et al.[40] used GMM and Kalman Filter for detecting and extracting the moving human and GRNN for collecting features of each frame and predicting the action . The sets used for evaluation are UCF Sports, UCF101, KTh and the accuracy achieved is 89.1.% 89.3 % and 96.3 % respectively.

Authors	Topic of Paper	Yearof blication	Dataset	Approach/Feature extraction	ummary	Accuracy
Wu et al.[41]	Action recognition by hidden temporal models	2014	Olympic , sports,HMDDB51	Spatio Temporal approach- HOG, MBH, HOF for features	Proposed a system that represents each class of action by hidden temporal model..[41]	84.3,47.1
Yuan et al.[42]	3D R Transform on STIP for	2013	KTH,UCF sports	Spatio Temporal	Proposed a global feature to	95.4
	Recognition of Action			approach- BoVW model	extact the complete geometrical location of points of interest .	
Singh et al.[43]	Multi- view recognition system for human activity based on multi- features for surveillance videos.[43]	2019	KTH	Stochastic Approach	In this a framework has been proposed that recognizes action on multiview bases.[43]	99.43
Baccouche et al.[45]	Sequential Deep Learning for Human Action Recognition	2011	KTH1,KTH2	Spatio Temporal/3D ConvNets	Proposed a fully automated model using RNN to classify	94.39,92.17

Subedar et al.[62]	Uncertainty-awareness Audio and visual Action Recognition with the help of Deep Bayesian Vibrational Inference[62]	2018	MiT	DNN based Work	Proposed a multimodal that is an uncertainty aware modal using Bayesian fusion framework[62]	81.2
Khan et al.	Hand-crafted and deep convolutional neural networks on features fusion and choosing strategy.To make a applicaton that is smart in human	2020	UCF sports,UCF11,Weiz mann	DNN based Work-Afusion of HOG feature with deep features.	Proposed a noval HAR system which have fusion of Histogram of Oriented Gradients and deep features and used SVM for classification	99.9,100,99.4
	action recognization.					
Liu et al. [46]	For recognition of human action by system Trust Gates are us	2016	Sbu,Berkeley MHAD, UT kinect	Extended already developed LSTM	Introduced new gating mechanism within LSTM for learning the goodness of input	93.3.,100, 97

	ed with Spatio Temporal LSTM				sequential data and effects are adjusted during update of LSTM memory .	

Table 5 Showing Different HRA systems developed in past 10 Years

5. FUTURE WORK

Human Action Recognition systems have been applied in many areas. It has been used in hospitals, parking areas, forstudying the environment, anomaly detection on airports, railway stations etc. But still more areas can be exported in future. There are many issues that still need some more attention like:

Surveillance Systems in Market Areas:

There are many commercial and public areas where systems have been installed like airports, railway stations, and markets but no much research has been done on shopping areas where changes of stealing goods and item is more. AI system can also help in analyzing customer behavior.

Real-Time Work

There is a need for more real-time based systems as most of the research that has been done is on recorded data withoutreal-time implementations.

Complex Activities

There are HAR systems that work very well on simple datasets but they fail when data is a little complex. So, there isa vast scope for research where multiple activities are present.

Prediction of Action

There are many areas where prediction of activity or action can be very helpful. These systems can understand the patter and can predict the future action in advance.

Working on Crowd

Classifying the activities of crowd is not an easy work. So, a lot of improvements can be done to enhance the results.

6. CONCLUSION

Action Recognition systems are very useful and have various applications in real life. These systems can solve our many problems related to security, prediction, surveillance, healthcare, etc. There are many areas where human life can be at risk. A lot of research has been done in past years. In this Survey, sincere efforts have been made to discuss all the techniques that evolved in the past 10 years. All the datasets available on HAR has also been included in the survey. Work done using these datasets is also included that can help the readers to understand which dataset and techniques can best work in their problem. In the last scope of research has been discussed that can help the upcomingresearchers.

REFERENCES

1. Recognition. In *BMVC* (pp. 1-12).
2. Quattoni, A., Wang, S., Morency, L. P., Collins, M., Darrell, T., & Csail, M. (2007). Hidden-state conditional random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(10), 1848-1852.
3. Park, S., & Aggarwal, J. K. (2004). A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia systems*, *10*(2), 164-179.
4. Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A Bayesian computer vision system for modeling human interactions. *IEEE transactions on pattern analysis and machine intelligence*, *22*(8), 831-843.
5. Prince, S. J. (2012). *Computer vision: models, learning, and inference*. Cambridge University Press.
6. Perronnin, F., & Dance, C. (2007, June). Fisher kernels on visual vocabularies for image categorization. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.
7. Wang, Y., & Mori, G. (2010, September). A discriminative latent model of object classes and attributes. In *European Conference on Computer Vision* (pp. 155-168). Springer, Berlin, Heidelberg.
8. Jayaraman, D., & Grauman, K. (2014). Zero-shot recognition with unreliable attributes. *Advances in neural information processing systems*, *27*.
9. Ji, S., Zhang, C., Xu, A., Shi, Y., & Duan, Y. (2018). 3D convolutional neural networks for crop classification with multi-temporal remote sensing images. *Remote Sensing*, *10*(1), 75.
10. Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015, June). Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-fourth international joint conference on artificial intelligence*.
11. Safaei M, Foroosh H (2017) Single image action recognition by predicting space-time saliency, pp 1–9
12. Banerjee, A., Singh, P. K., & Sarkar, R. (2020). Fuzzy Integral-Based CNN Classifier Fusion for 3D Skeleton Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(6), 2206-2216.
13. Aubry, S., Laraba, S., Tilmanne, J., & Dutoit, T. (2019). Action recognition based on 2D skeletons extracted from RGB videos. In *MATEC Web of Conferences* (Vol. 277, p. 02034). EDP Sciences.
14. Khan, M. A., Javed, K., Khan, S. A., Saba, T., Habib, U., Khan, J. A., & Abbasi, A. A. (2020). Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, 1-27.
15. Dai, C., Liu, X., & Lai, J. (2020). Human action recognition using two-stream attention based LSTM networks. *Applied soft computing*, *86*, 105820.
16. Li, W., Wen, L., Chang, M. C., Nam Lim, S., & Lyu, S. (2017). Adaptive RNN tree for large-scale human action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 1444-1452).
17. Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., & Van Gool, L. (2019). stagNet: an attentive semantic RNN for group activity and individual action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 549-565.
18. Gao, X., Luo, H., Wang, Q., Zhao, F., Ye, L., & Zhang, Y. (2019). A human activity recognition algorithm based on stacking denoising autoencoder and lightGBM. *Sensors*, *19*(4), 947.

19. Almaslukh, B., AlMuhtadi, J., & Artoli, A. (2017). An effective deep autoencoder approach for online smartphone-based human activity recognition. *Int. J. Comput. Sci. Netw. Secur*, *17*(4), 160-165.
20. Abdellaoui, M., & Douik, A. (2020). Human Action Recognition in Video Sequences Using Deep Belief Networks. *Traitement du Signal*, *37*(1).
21. Foggia, P., Saggese, A., Strisciuglio, N., & Vento, M. (2014, August). Exploiting the deep learning paradigm for recognizing human actions. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 93-98). IEEE.
22. Kanjo, E., Younis, E. M., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, *49*, 46-56.
23. Jaouedi, N., Boujnah, N., & Bouhlel, M. S. (2020). A new hybrid deep learning model for human action recognition. *Journal of King Saud University-Computer and Information Sciences*, *32*(4), 447-453.
24. Wu, J., Hu, D., & Chen, F. (2014). Action recognition by hidden temporal models. *The Visual Computer*, *30*(12), 1395-1404.
25. Yuan, C., Li, X., Hu, W., Ling, H., & Maybank, S. (2013). 3D R transform on spatio-temporal interest points for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 724-730).
26. Singh, R., Kushwaha, A. K. S., & Srivastava, R. (2019). Multi-view recognition system for human activity based on multiple features for video surveillance system. *Multimedia Tools and Applications*, *78*(12), 17165-17196.
27. Chen, C. Y., & Grauman, K. (2016). Efficient activity detection in untrimmed video with max-subgraph search. *IEEE transactions on pattern analysis and machine intelligence*, *39*(5), 908-921.
28. Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011, November). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding* (pp. 29-39). Springer, Berlin, Heidelberg.
29. Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016, October). Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision* (pp. 816-833). Springer, Cham.
30. Kanjo, E., Younis, E. M., & Ang, C. S. (2019). Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion*, *49*, 46-56.
31. Ijjina, E. P. (2016). Classification of human actions using pose-based features and stacked auto encoder. *Pattern Recognition Letters*, *83*, 268- 277.
32. Nicolaou, M. A., Pavlovic, V., and Pantic, M. (2014). Dynamic probabilistic CCA for analysis of affective behavior and fusion of continuous annotations. IEEE Trans. Pattern Anal. Mach. Intell. 36, 1299–1311. doi:10.1109/TPAMI.2014.16
33. Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2014b). "Classifying behavioral attributes using conditional random fields," in Proc. 8th Hellenic Conference on Artificial Intelligence, Lecture Notes in Computer Science, Vol. 8445 (Ioannina), 95–104.
34. Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. (2014). Learning multimodal latent attributes. IEEE Trans. Pattern Anal. Mach. Intell. 36, 303–316. doi:10.1109/TPAMI.2013.128
35. Gan, C., Wang, N., Yang, Y., Yeung, D. Y., and Hauptmann, A. G. (2015). "DevNet: a deep event network for multimedia event detection and evidence recounting," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 2568 – 2577.

36. Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. IEEE Trans. Pattern Anal. Mach. Intell. 29, 2247–2253. doi:10.1109/TPAMI.2007.70711

37. Fernando, B., Gavves, E., Oramas, J. M., Ghodrati, A., and Tuytelaars, T. (2015). "Modeling video evolution for action recognition," in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Boston, MA), 5378–5387.

38. Holte, M. B., Tran, C., Trivedi, M. M., and Moeslund, T. B. (2012b). Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. IEEE J. Sel. Top. Signal Process. 6, 538–552. doi:10.1109/JSTSP.2012.2196975

39. Ronao, C. A., & Cho, S. B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. Expert systems with applications, 59, 235-244.

40. https://www.di.ens.fr/willow/research/stillactions

[58] http:// host. robots. ox. ac. uk/ pascal/ VOC

[59] http:// vision. stanf ord. edu/ Datas ets/ 40act ions. Html

[60] https:// www. di. ens. fr/ willow/ resea rch/ still actio ns/

[61] https://www.crcv.ucf.edu/research/data-sets/ucf101/

[62] Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O., & Huang, J. (2019). Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 6301-6310).